The Rationality of Fair Division as an Equilibrium Outcome in the Ultimatum Game

TAYLOR W. BULEY

Economic man is self-interested and rational. But in practice, humans often act irrationally and sacrifice their own self-interest for the benefit or detriment of another's outcome. Rooted in the same fundamental economic principles, is formal game theory, then, a wash? This paper sets out to bring together evidence from recent discussion in behavioral and evolutionary game theory in an attempt to rationalize fair division as a short run equilibrium strategy played by self-interested and rational human beings in the ultimatum game.

Introduction

Is formal game theory supposed to describe actual behavior or not? It's a seemingly simple question that game theory has successfully avoided for a surprising amount of time.

Imagine the following situation, which is called the 'ultimatum game' and is frequently employed by game theorists as a way to model human behavior. The rules are as follows: one bargainer (henceforth known as the 'first-mover') makes a proposal of how to divide a windfall of a certain amount with another bargainer (the 'second-mover'), who has the opportunity to accept or reject the proposed division. If the second-mover accepts, each bargainer earns the amount proposed for him by the first-mover, and if the second bargainer rejects, then each bargainer earns zero.

This stylized brand of negotiation was first studied by experimental economists, and their paper surprised traditional economists with the following solution: game theoretic analysis

favors an unequal split in equilibrium, predicting that the first-mover will walk away with everything in the end, and the second-mover with nothing. Furthermore, the solution is a subgame perfect equilibrium, meaning that at that particular 'node' in the game neither player could do any better by switching strategies (Güth *et al*. pp. 367-388).

Experimental evidence presents a compelling argument that in any kind of multistage game, players anticipate their opponents' future actions. Assuming that he wants to maximize his monetary payoff, economists take this to mean that the second-mover in the game will accept any positive offer that the first-mover offers her under the simple assumption that any player would rather accept the offer and earn whatever she can, rather than reject it and earn zero. Eventually realizing that the second-mover will accept any non-negative offer that he offers her, in later rounds the first-mover will inevitably offer just the minimal amount necessary to keep the offer non-negative.

Formal game theory makes no claim over the "fairness" of this equilibrium –

assuming self-interest and rationality, the second-mover would do the same thing given first mover advantage. But in practice, human behavior frequently deviates from the game theoretic predictions (Halpern, "Elements" pp. 835-868). Ochs and Roth (pp. 355-384), among others, have noted that a notion of "fairness" might be influencing players' behavior. How can we incorporate this idea into traditional game theory?

Fair Division in the Long Run

It is well known in economic literature that subjects do not anticipate future actions as they should according to game theoretic models (Harsanyi and Reinhard, pp. 80-106). Nevertheless, the fact that experiments in bargaining strategy consistently yield out-of-equilibrium results is bothersome to game theorists.

How does formal game theory cope with the concept of fair division? According to evolutionary models of behavior, randomly matched individuals playing the ultimatum game over the long run will typically observe one division almost all of the time. When all individuals in the population adopt the same strategy, this stochastically stable division yields close payoffs to those of the subgame perfect equilibrium (Young , pp. 145-68). Thus, we

can argue that fair division as equilibrium in the long run is at least theoretically possible, given the right conditions.

Brian Skyrms goes beyond this claim to assert that not just can the right starting conditions carry the "demand half" strategy to fixation according to discrete replicator dynamics, but the demand half strategy is in fact the only evolutionarily stable strategy (i.e. the only strategy robust to 'mutant' innovators if the whole population were to adopt the strategy). In short, fair division isn't only possible – it's also "more likely than not" (Skyrms, 2003, pp.17-29).

Skyrms, along with Jason Alexander, ran a large simulation in a population of 10,000 plotted on a square lattice, starting repeatedly at randomly chosen starting points to explore how "inevitable" the fair division outcome is and measuring the basins of attraction for the various polymorphic pitfalls on the way. Sure enough, fair division went to fixation in more than ninety-nine point five percent of the trials. The cases where it did not were all cases where the initial population contained fewer than seventeen players playing the demand-half strategy.

Skyrms and Alexander's experiment shows that there are contagious dynamics of equal division when bargaining with neighbors. Bargaining with strangers in the ultimatum game leads to fair division from a randomly chosen starting point about 60 percent of the time. As soon as a small group of players playing the demand-half strategy form, justice is contagious and fair division takes over. In the "ultra-ultra long run," fair division is the inevitable conclusion (Skyrms and Alexander, pp. 588).

Evolutionary game theory explains how "demand half" is a stochastically stable outcome in the ultra long run; however, as Skyrms points out, long expected waiting times call into question the explanatory significance of evolutionary analysis. We know that demanding half bodes well as a strategy in the long run, but what makes it better than the dominant strategy of demanding the total sum in the short run? As Camerer opines, there is no triumph for formal game theory until it can explain the behavior in early rounds that produces the fair division equilibrium in later rounds (Camerer, pp.167-88).

Fair Division in the Short Run

We know that "demand half" is a stochastically stable strategy in the long run, but can behavioral game theory explain how application of the fair division strategy can yield a subgame perfect equilibrium in the short run without the aid of any unnecessary psychological handwaving?

Lopomo and Ok offer a model capable of accommodating observed differences in experimental bargaining games and game theoretic predictions while still providing a rational theory of fair outcomes, which they achieve by rigorously modeling a concept they call "fear of rejection" (Lopomo and Ok, pp. 263-83). The notion presumes that people derive negative utility from rejection, and thus yield positive utility by avoiding it. The main contribution of the model is that it allows for the negative interdependence of preferences – that is, it allows for "altruism," the idea that some people derive positive psychological utility by helping others, and that if large enough, this utility can dominate material preferences in utility maximization.

Nevertheless, there are at least three major concerns with Lopomo and Ok's attempt at modeling bargaining behavior in the ultimatum game. The first is that the notion of "fear of rejection" depends on negatively interdependent preferences, which are not part of the game but rather implicit in the preferences and beliefs of the players. Interdependence of preferences keeps game theorists from modeling players as truly independent of each other, undermining the fundamental economic assumption that humans are self-interested beings. The second concern is over the model's asymmetric dependence on negatively interdependent preferences (i.e. altruism) without considering the contra positive situation (i.e. positively interdependent preferences), commonly known as 'spite' and commonly neglected in the discussion of fairness norms in formal game theory thus far. Finally, the assumption that players are rational allows for players to think (i.e. assign a positive probability to the event) that his opponent may care about his relative share of the whole – which calls into contests the usefulness of the model altogether.

An earlier model developed by Matthew Rabin proposes a formal way to explain fair division in the short run, using rational choice economics, without the need for psychological handwaving. Rabin's framework enables us to incorporate emotions into a broad number of economic models, developing a theoretic solution he calls a "fairness equilibrium" that

incorporates psychological evidence that (1) people are willing to sacrifice their own material well-being to help those who are being kind, (2) people are willing to sacrifice their own material well-being to punish those who are being unkind; and (3) fairness influences behavior the most when material stakes are low without having to rule out self-interest and rationality as fundamental assumptions (Rabin pp. 1281-1302). Modeling emotions like altruism and spite allows us to begin to understand their economic and social implications more generally.

Rabin divides utility into two inputs, material and psychological, and it can be both positive and negative. Intuitively, if a player thinks his opponent is going to act fairly toward him, he is more likely to act fairly in return. When both act fairly, both derive positive psychological utility. Acting fairly toward a partner who acts unfairly yields negative psychological utility, but a player can also produce utility by acting selfishly against someone who acts selfishly toward him.

Rabin's model does a good job filling the gap in coverage produced in cases where players have positively interdependent preferences by Lopomo and Ok's model. It also formally allows players the capacity to assign a positive probability to the event that his opponent cares about his relative share of the whole. Unfortunately, Rabin's model alone it still is not enough to model our experimental results from the ultimatum game.

If it were true that the fairness equilibrium was a complete model, we should be able to extend any complete model to incorporate evolutionary game theoretic results. However, it is not the case that Rabin's model can explain Brian Skryms' finding that fairness is a stochastically stable strategy that reaches fixation in the long run "more likely than not." This is not yet the case with Rabin's model. His model assumes that the utility derived from material payoffs in the ultimatum game monotonically increases along with the stakes, but the utility derived from the fairness remains constant as the stakes increase. Thus, in the long run, players' strategies are dominated by material outcome and fairness becomes indistinguishable from the subgame perfect equilibrium.

In a response to this problem in Rabin's "fairness equilibrium," William Robert Nelson, Jr. points out a small adjustment in Rabin's model that can make it infinitely more useful in explaining the kind of behavior in early rounds that produces the fair division equilibrium in later rounds. Nelson points out that by treating fairness as a standard economic good,

instead of as a normal good, you can allow for the theoretical situation in a game where the payoffs are *very very large* and concerns over fairness dominate a player's utility derived from material goods. If you incorporate the idea of diminishing marginal utility of wealth, fairness considerations may, at *very very high* stakes, dominate players' material considerations (1180-1183). Simply put: the richer you get, the less you have to worry about material needs and the more you can afford to focus on psychological needs. You can also see this effect in action in today's society, where corporate philanthropy is a virile business strategy and the world's richest man is also its biggest philanthropist (Conlin and Hempel).

With the Nelsen addendum, Rabin's model of "fairness equilibrium" is of great aid to the game theorist in explaining how behavior in the early rounds leads to stochastically stable fair division equilibrium in the limit. The U-shaped utility curve of fairness yields fair division in the limit as concerns for fairness dominate material wealth. The model can account for both altruism and spite, which are widely observed in the ultimatum and yet absent from in formal game theoretic model of human behavior.

Nevertheless, there is one final concern with this amended version of Matthew Rabin's model: like with Lopomo and Ok's attempt at modeling bargaining behavior in the ultimatum game, the concept of "fairness equilibrium" depends on negatively and positively interdependent preferences, which are not part of the game but rather implicit in the preferences and beliefs of the players. Interdependence of preferences keeps game theorists from being able to model players as truly independent of each other – which undermine the economic assumptions behind formal game theory. Any complete attempt at modeling bargaining behavior would have to be able to incorporate independence of preferences.

Bringing It All Together

Now that we can rationalize fair division in the short run using utility functions adjusted to reflect the psychological utility a player derives from achieving a fair division, can we use predictions made in evolutionary game theory to pull expected utility up by the bootstraps and accommodate the short run experimental results?

Predictions in evolutionary game theory give us a leg up in this endeavor: from models of neighborhoods interaction over the long run, we know that you'll reach fair division quicker bargaining with neighbors than you will when bargaining with strangers. A study done

by Jennifer Halpern confirms this prediction in the short run. Halpern examined how friendship alters expectations for pricing strategies in bargaining games and provides the evidence we need to confirm the equilibrium prediction of fair division in evolutionary game theory. Subjects were told that they were participating in a study designed to explore how people "like you" make pricing decisions. The experimenter randomly assigned the subjects to roles as buyer or seller and gave them a list of several commodities familiar to them. They were asked to imagine themselves involved in a transaction with someone of the same biological sex who was described as either a friend or someone that they did not know who is also a fellow student ("Effects" 64-68).

The following hypothetical commodities with their typical price ranges were included: concert tickets ($10-$26); dictionary ($5-$20); calculator ($20-$50); telephone ($50-$100); and a television ($175-$200). After reading each description, participants were asked to indicate how much they would offer if assigned to the buyer role, or accept for the item if in the seller role. The participants were reminded that negotiation was not possible.

For all five commodities, the friend was willing to accept an amount below the midpoint of the commodity's typical price range, so friends were not just merely "splitting the difference" by selecting an obvious price point midway between the high and low end of the price range recognized by both parties, as is often suggested early papers on bargaining behavior. Halpern observed that that friend-sellers were willing to accept less than stranger-sellers, and friend-buyers were willing to pay more than stranger-buyers. Participants consistently offered more to hypothetical-friends then hypothetical strangers for all commodities. A similar hypothesis was confirmed about the effects of friendship on pricing expectation: participants consistently asked for more for all commodities from hypothetical strangers than from hypothetical friends.

The study also found that while stranger-buyers offered significantly less than stranger-sellers, friend-buyers and friend sellers *agreed* on price. It is not surprising that strangers disagreed about pricing (this gap in utility functions is what enables bargaining to occur); rather what is surprising is that these participants made these pricing decisions – which called for friends to agree and strangers to disagree – completely independently of one another, without the cues available in negotiation and with no signaling system in place.

Halpern's study strongly suggests that there is a tendency to alter one's expectations for an exchange based on anticipated future interaction (i.e. friendship). Unlike the model proposed by Lopomo and Ok which relied upon the outside concept of "fear of rejection" and did not allow us to cover situations where spite dominates behavior, the interaction between buyers' and sellers' roles in this study clearly shows that friendship is indeed a factor mediating price expectations.

What do these findings mean in context of evolutionary replicator dynamics and evolutionary game theory? As Skyrms showed in his neighborhood models of interaction, players playing the "demand-half" strategy do better in groups of neighbors playing the same strategy. Is the behavioral game theoretic model that we have just developed able to incorporate these payoffs?

The idea of social norms for friendship allows us to understand Halpern's experimental evidence that humans follow bargaining scripts that distinguish between friends and neighbors. A seller discounts a commodity for a friend because she trust the friend to return the favor someday; the seller also expects there to be a "someday" in the future during which they will interact and wants to produce goodwill and maintain the harmonious interactions they have exchanged. Halpern calls this the notion of "scripts for friendship" – norms for fairness may call for us to play the "demand half" strategy with friends, but our scripts for dealing with strangers make no such considerations.

The concept of anticipated future interaction is helpful, but not necessary, to understanding friendship scripts. For example, if you are chosen as a godfather for a dying friend, you are probably likely to still honor the obligation to raise the child even though you know there will not be any future opportunities for him to return the favor. Similarly, the notion of trust is a necessary precursor to the "demand half" strategy but not sufficient enough to cue friendship scripts when dealing with strangers. This sheds some light on the Skyrms finding that the key to the contagion of a strategy is interaction along the edges of the "patch" as mapped on a lattice. You can trust that a department store's reputation for being fair to its consumers, but that doesn't mean that you will offer to pay more for their merchandise on account of that trust.

Fairness equilibrium and anticipated future interaction help explain the substantial frequency of rejected offers in the ultimatum game, a task which formal game theory has yet

to do. These rejected divisions, which result in Pareto-inefficient outcomes, do not occur in the traditional game theoretic model, which assumes that as utility maximizers players would accept any proposed division rather than reject it and gain nothing from the transaction. In the laboratory, however, the probability that an offer will be rejected is inversely related to the size of the offer within country irrespective of cultural norms of fairness. Furthermore, higher disagreement rates are not observed in countries where lower offers are observed and the probability that an offer is rejected is actually lower in countries where a lower offer is observed (Roth *et al.* 1068-95).

Conclusion

Traditional game theory, deeply rooted in economics, has had similar troubles asserting its relevance beyond the theoretical as a descriptive approach to modeling human behavior. Like economists, game theorists assume that humans everywhere deploy the same cognitive machinery for making economic decisions, and consequentially would behave similarly when faced with comparable situations. To the contrary, experimental evidence strongly suggests that social norms of the perception of fairness play a strong role in bargaining behavior, and suggest that these norms may vary across cultures (Henrich 973-979). Norms for fairness – which include both altruism and spite – helps extend traditional game theory to incorporate how human behavior differs from traditional game theoretical prediction. Matthew Rabin's model of "fairness equilibrium," combined with the idea of anticipated future interaction and the Nelsen addendum of a U-shaped curve of the psychological utility of fairness, rationalize the fair division equilibrium in the limit, as predicted by evolutionary game theory, without having to abandon the fundamental assumption that humans are self-interested and rational beings.

REFERENCES AND CITATIONS

Camerer, Colin F. "Progress in Behavioral Game Theory." *Journal of Economic Perspectives*; Fall 1997, Vol. 11 Issue 4, pp. 167-188.

Halpern, Jennifer J. 1991. The effects of friendship on bargaining: Experimental studies of personal business transactions. *Academy of Management Best Papers Proceedings*, edited by J. L. Wall and L. R. Jauch, Las Vegas, NV: Academy of Management, pp. 64-68.

Halpern, Jennifer J. "Elements of a Script for Friendship in Transactions." *The Journal of Conflict Resolution*, Vol. 41, No. 6., December 1997, pp. 835-868.

Harsanyi, John C. and Reinhard Selten. "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information." *Management Science*, Vol. 18, No. 5, Theory Series, Part 2, Game Theory and Gaming, January 1972, pp. 80-106.

Henrich, Joseph. "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga of the Peruvian Amazon." *American Economic Review*; Sep2000, Vol. 90 Issue 4, pp. 973-979.

Lopomo, Giuseppe and Efe A. Ok. "Bargaining, Interdependence, and the Rationality of Fair Division." *The RAND Journal of Economics*, Vol. 32, No. 2. Summer 2001, pp. 263-283.

Ochs, Jack and Alvin E. Roth. "An Experimental Study of Sequential Bargaining." *The American Economic Review,* Vol. 79, No. 3. June 1989, pp. 355-384.

Rabin, Matthew. "Incorporating Fairness into Game Theory and Economics." *The American Economic Review*, Vol. 83, No. 5., December 1993, pp. 1281-1302.

Roth, Alvin E.; Prasnikar, Vesna; Okuno-Fujiwara, Masahiro and Zamir, Shmuel. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, December 1991, 57(5), pp. 1068-95.

Skyrms, Brian and Jason Alexander. "Bargaining with neighbors: Is justice contagious?" *Journal of Philosophy;* November 1999, Vol. 96 Issue 11, p 588.

Skyrms, Brian. "The Stag Hunt and the Evolution of Social Structure." *Cambridge University Press*, Cambridge, U.K., 2003, pp. 17-29.

Conlin, Michelle and Jessi Hempel. "The Top Givers." *BusinessWeek*, 1 December 2003.

Werner Güth, Rolf Schmittberger and Bernd Schwarze. "An experimental analysis of ultimatum bargaining." *Journal of Economic Behavior & Organization*, Volume 3, Issue 4, December 1982, pp. 367-388.

William Robert Nelson, Jr. "Incorporating Fairness into Game Theory and Economics: Comment." *The American Economic Review*, Vol. 91, No. 4. September 2001, pp. 1180-1183.

Young H. P. "An Evolutionary Model of Bargaining." *Journal of Economic Theory*, Volume 59, Issue 1, February 1993, pp. 145-168.